Domain-Independent Detection, Extraction, and Labeling of Atomic Events

Elena Filatova and **Vasileios Hatzivassiloglou**
Department of Computer Science
Columbia University
New York, NY 10027, USA
{filatova, vh}@cs.columbia.edu

Abstract

The notion of an "*event*" has been widely used in the computational linguistics literature as well as in information retrieval and various NLP applications, although with significant variance in what exactly an event is. We describe an empirical study aimed at developing an operational definition of an event at the *atomic* (sentence or predicate) level, and use our observations to create a system for detecting and prioritizing the atomic events described in a collection of texts. We report results from testing our system on several sets of related texts, including human assessments of the system's output and a comparison with information extraction techniques. We discuss how event detection at this level can be used for indexing, summarization, and question-answering.

1 Introduction

> "The world changes, things happen, time passes. We live in a world where events, both important and mundane, define and demarcate our lives."[1]

What is an event? It seems that, like life, this is a term that is hard to define precisely, though easy to understand at the intuitive level.

WordNet[2] (Miller et al., 1990) defines an event broadly as "something that happens at a given place and time". Linguists that have worked on the underlying semantic structure of events and their realization in text propose more complicated definitions involving telicity, time, and external world conditions; for example, Chung and Timberlake (1985) state that "an event can be defined in terms of three components: a predicate; an interval of time on which the predicate occurs, ...; and a situation or set of conditions under which the predicate occurs, ..." Pustejovsky (2000) argues for a semantic theory of events that models persistence as well as change and is grounded on the notion of predicate opposition between objects and properties. He notes that "lexical semanticists must look outward from the verb to the sentence in order to characterize the effects of a verb's event structure; and logical semanticists must look inward from the sentence to the verb to represent semantic facts that depend on event-related properties of particular verbs".

Recent work in information retrieval within the TDT framework has taken event to mean essentially "narrowly defined topic for search" (Allan et al., 1998; Yang et al., 1999). On the other hand, within the information extraction community, an event represents a relationship between participants, times, and places; the Message Understanding Conference (Marsh and Perzanowski, 1997) defines one of its tasks as "extract prespecified event information and relate the event information to particular organization, person, or artifact entities involved in the event."

Not only is the exact meaning of events in dispute, but also the extent of an event's realization in text. As it has been mentioned above, most linguists associate events with the tensed matrix verb of a sentence or simple clause, and by extension with that sentence or clause. However, events can be expressed with a single noun phrase such as "war" and "strike" (Pustejovsky, 2000), and sentences can describe multiple events in separate simple clauses (Filatova and Hovy, 2001). In the Topic Detection and Tracking (TDT) framework (Allan et al., 1998; Yang et al., 1999), the ongoing DARPA/NIST effort on text categorization and clustering, events are represented as sets of related documents. Finally, MUC's events are represented as predefined templates, with attributes corresponding to participants, location, and time filled in.

Detecting events automatically and obtaining a semantic representation for them would be equivalent to creating a "Who did what to whom when and where?" interpretation of the text. Such an interpretation, as we argue in Section 8, would offer new venues for research in text indexing, summarization, visualization of information, and question answering. In this paper, we explore an automatic method for creating such an interpretation at the atomic (sentence or predicate) level, extracting and representing multiple "small" events rather than only the ones at the highest level. We aim at small text pieces, unlike the TDT/information retrieval approach, but we utilize similarities between related texts to determine which atomic events and relationships are specific to a broader event or topic. We draw from linguistic analysis and theory, but present an operational method for analyzing arbitrary texts. We trade off "understanding" of the text to the extent that information extraction achieves it for the generality of operating without the constraints of a specific domain. Our representation is necessarily weaker than the predefined templates used in IE tasks, but covers many types of events which are labeled with appropriate verbs and nouns selected from the texts.

In Section 3, we discuss a study of events in newswire articles which explored how well people can detect (and agree on) events at the atomic level. We then present

[1] Frank P. Coyle, in *Ubiquity: an ACM IT Magazine and Forum*, Volume 4, Issue 4, March 18 - 24, 2003.

[2] http://www.cogsci.princeton.edu/~wn/.

an automated system that relies in part on the findings of this study to detect relationships between entities of certain types (by default named entities), isolate the likely events, collate events involving the same entities, and label the combined event with nouns and verbs. Sections 5–7 report sample results and an evaluation of the output of our system. We conclude by discussing how we plan to use this output for a variety of NLP tasks.

2 Related Work

As noted in the introduction, events and their semantic structure have been analyzed by several linguists, who have looked at semantic constraints in sentences to distinguish between events, extended events, and states; see for example (Chung and Timberlake, 1985; Bach, 1986; Pustejovsky, 2000). Often in such type of research event analysis is centered on properties of the verb, and verbs are classified according to their relationships with event classes (Levin, 1993). From a computational perspective, discourse analysis has relied on (often implicitly defined) events; for example, McCoy and Strube (1999) investigated time intervals that can be assigned to events (atomic events occurred at a singe point in time versus repeated atomic events, extended atomic events or states that occurred over a span of time) to generate pronouns. In this work simple clauses were taken as the text region for events. Siegel and McKeown (2000) have proposed automatic methods for classifying verbs according to whether they can signal events and processes (stativity and completeness). Filatova and Hovy (2001) built a system for assigning time stamps to the event clauses and recognized the problem of locating the extent of events in text when they needed to determine the scope of each detected time expression.

The work most commonly referred to as event detection is that originating from the Topic Detection and Tracking (TDT) research effort sponsored by DARPA. An important contribution of that research program is the recognition of the distinction between an event and a topic; however, this distinction is made principally on the basis of specificity and targeted information retrieval rather than linguistic properties of the retrieved units. As Yang *et al.* (1999) note, "[the] USAir-427 crash is an event but not a topic, and 'airplane accidents' is a topic but not an event". In practice, the TDT data sets included "events" with a widely varying scope, from "Comet into Jupiter" to "Oklahoma City bombing", and never aimed at extracting information at less than the document level or structuring that information with semantic role annotation (although some current directions in TDT, such as new information detection (Allan et al., 2001), operate on text passages smaller than the entire document).

Systems participating in the Scenario Template task of the Message Understanding Conference (MUC) competitions (from 1992 to 1998) use information extracted and inferred from a text to fill in the appropriate fields in predefined templates corresponding to the domain of the text. Since the domain is given and the semantics for each field is known, systems can achieve fairly high performance (up to 50%–60% recall and precision) on a useful text understanding task. In many ways, the approach taken in MUC is similar to ours, in that we also aim to retrieve relationships between participants, times, and locations in events, and label the extracted events to reflect those relationships. However, the MUC systems suffer from two drawbacks: First, the fixed templates preclude detecting multiple events of different types, or of types not anticipated during system design.[3] Second, they are heavily dependent on the domain, which requires a lot of time to create accurate templates defining possible events for that particular domain, and even more effort in adapting the system to the sublanguage and knowledge model of that domain.

3 A Study of Event Annotation

We conducted a first study of text annotation for event information by asking a number of computer science graduate students (mostly in computational linguistics) to mark text passages that describe events in news stories. We deliberately provided no definition of *event* for this study, to see if the respondents would naturally converge to an operational definition (as evidenced by high agreement on what they marked). The annotators were given 13 news articles randomly selected from the DUC-2001 (Document Understanding Conference) corpus. The texts varied in length from 15 to 60 sentences. Five of the thirteen texts were each annotated by two participants in the study. In addition to checking for agreement between the annotators and anecdotal evidence of the difficulty or ease with which they could label events, our study had two further aims: To determine what text ranges, in the absence of instructions on the length of what they should mark, people tend to favor as the appropriate text parts describing a single event; and to gather evidence of features that occur with high frequency in the marked passages and could be automatically extracted by an automated system simulating the human annotators.

3.1 Agreement

We noticed substantial disagreement between annotators on what should be marked as an event. Recall that our instructions for this experiment only asked them to find the important events in a given text, providing no definition. In that context, people often disagreed on whether a given passage should be marked as an event description or not. Since our annotation instructions left unspecified the length of event descriptions, a basic text unit that could be marked or unmarked is not defined either and therefore it is hard to quantitatively measure the agreement between the annotators. Nevertheless, we made several qualitative observations on the basis of repeated patterns of disagreement:

- In some cases, an important part of the text that nevertheless represented a continuation of a state was marked as an event, for instance

  We have no quarrel with the people of Iraq.

[3] For example, as we show in Section 6, they can detect a *kidnapping* event but not the victim's *release* as a separate event.

- Related events that occurred sequentially in text were sometimes grouped in one marked text region, as in

  > *The Soviet Union said today it had sent an envoy to the Middle East on a series of stops to include Baghdad. Soviet officials also said Soviet women, children, and invalids would be allowed to leave Iraq.*

- Annotators often disagreed consistently on the marking of specific subtypes of events. One such subtype that is common in news stories is an *utterance* event, i.e., an event where the protagonist says, announces, or describes something. The act of the utterance is an event according to most definitions, but depending on whether the thing being said is also an event and how important that thing is, annotators marked the entire sentence as an event or non-event.

- Further, analysis of the responses showed that often a single annotator was not consistent in their own assessments across similar types of text passages. For example, one of the annotators marked the passage

  > *The British Foreign Office said today conditions in Kuwait appear to be deteriorating.*

  as an event, but did not mark the similar passage

  > *The predominantly Moslem nation of Bangladesh said today its troops would join multinational forces in Saudi Arabia.*

  The annotators' reduced ability to distinguish between utterance-type events and other events is compounded here.

3.2 Length of Marked Text Passages

While the annotators disagreed on what text pieces to select as event descriptions, they exhibited more agreement on how long these pieces should be. Out of 190 text regions marked as events, 46 (24%) consisted of one clause within a longer sentence, 22 (11%) of one sentence minus one short prepositional phrase, 95 (50%) of exactly one sentence, and 27 (14%) of multiple sentences.

According to this analysis, the simple clause is really the minimal unit representing atomic events (noun phrases such as *war* or *earthquake* were never marked as events). However, twice as many full sentences as simple clauses were marked as events, and an additional 11% of the marked regions were almost full sentences. We therefore conclude that full sentences appear to provide the most reasonable scope for locating atomic events.

3.3 Text Features in Marked Passages

We analyzed the passages marked as event descriptions looking for text features that could be included in an automated event detection system. Naturally, the verb itself often provides important information (via tense, aspect, and lexical properties) about the event status of a clause or sentence. In addition, the following features are correlated

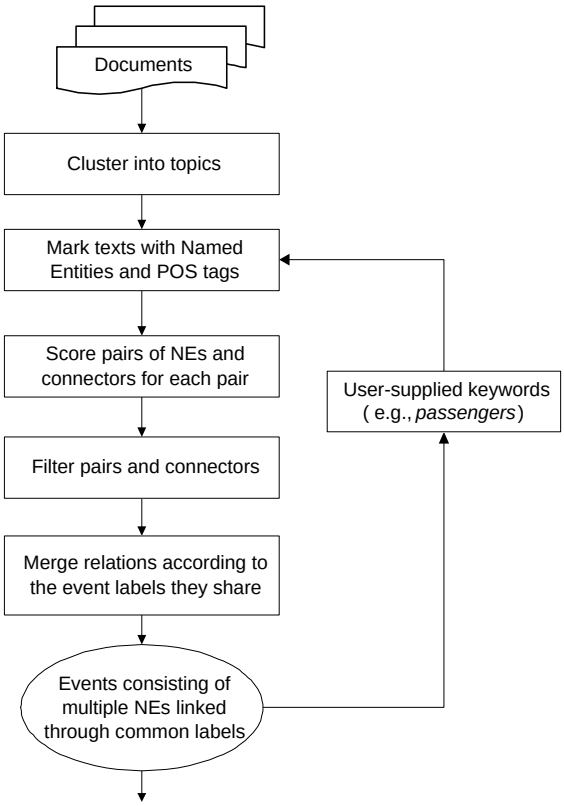

Figure 1: Outline of atomic event detection system.

with the presence of events: **Proper nouns** occur more often within event regions, possibly because they denote the participants in events. In contrast, **pronouns** are less likely to occur in event regions than in non-events. As expected, the presence of **time phrases** increases the likelihood of a text region being marked as an event description. **Cardinal numbers** were another lexical class strongly associated with events. This can be attributed to the fact that numbers are often given when new important information is presented; they condense information and typically accompany factual rather than subjective sentences, which are more likely to be associated with event descriptions.

4 Detecting and Labeling Events

Drawing from our event annotation study, we decided on an algorithm for detecting, extracting, and labeling events that is based on the features that seemed more strongly correlated with event regions. The outline of our approach is shown in Figure 1. We anchor events on major elements representing participants (proper nouns for people and organizations), locations (again typically proper nouns), and time information. All these major elements can be retrieved with a named entity tagger; we use BBN's IdentiFinder (Bikel et al., 1999), and we expect at least two such major elements in a sentence to consider extracting an event. We take the sentence as the scope of an event, and include in the extracted elements cardinal numbers (and the nouns they modify) as likely to be important in the event. Our algorithm ignores sentences that contain one named entity

(or none). Otherwise, we extract all the possible pairs (preserving the order of named entities) and all the words that are in between each pair of named entities. After extracting this information from the entire text (or set of related texts, see below), for each pair of named entities (a *relation*) we calculate how many times it occurs, irrespective of the in-between words (the *connectors*). For each connector we calculate how many times this connector is used in the extracted relation.

Our hypothesis is that if named entities are often mentioned together, these named entities are strongly related to each other within the topic from which the relation was extracted. Although our method can be applied to a single text (which by itself assures some topical coherence), we have found it beneficial to extract events from sets of related articles. Such sets can be created by clustering texts according to topical similarity, or as the output of an information retrieval search on a given topic.

We create a list of connected named entity pairs in decreasing frequency order, and a list of connectors between such pairs also in decreasing frequency order. We first filter the connector list, by keeping only verbs or nouns of frequency above a threshold; nouns must also be hyponyms of *event* or *activity* in WordNet. We use Charniak's (2000) statistical parser to obtain part of speech information. Finally, we eliminate from our candidate event list those pairs that are no longer supported with a significant number of connectors or are not among the top n events (both of these parameters are adjustable and are determined empirically).

We then examine the graph of connections induced by the surviving pairs. For each two edges in that graph with a common endpoint (e.g., (A, B) and (A, C)), we examine if their list of connectors is substantially similar. We consider two such lists substantially similar if one contains at least 75% of the elements in the other. When that condition applies, we merge the two candidate events into one link between A and a new element $\{B, C\}$ (i.e., we consider B and C identical for the purpose of their relationship to A), and add the scores of the two original events to obtain the score of the composite event. This pushes together similar named entities (for example, alternate spellings or two alternate descriptions for a person or location), reducing the redundancy of the extracted events without using any explicit knowledge about such relationships in the real world.

The relative order of extracted events is further modified by two additional factors: we prioritize pairs that occur in multiple documents within a set of related documents, and we reduce the importance of pairs that occur frequently in a large text collection. These steps highlight events that are *topic-specific*; pairs of entities that are linked irrespective of the specific events described in the set of texts being analyzed (e.g., Bush and Cheney) will thus be pushed further down in the list.

5 System Output

In this section we present and comment on sample system output. We ran our system on a subset of the topics provided by the Topic Detection and Tracking Phase 2 research effort (Fiscus et al., 1999). The topics consist of articles or transcripts from newswire, television, and radio (the New York Times, Associated Press, CNN Headline News, ABC World News Tonight, PRI The World, and Voice of America English News Service). We used 70 of the 100 topics, those containing more than 5 but less than 500 texts. Since human annotators created these topical clusters in a NIST-sponsored effort, we can be assured of a certain level of coherence in each topic. In this manner, we can concentrate on the benefits or shortcomings of our algorithm rather than on issues related to the retrieval of on-topic texts.

THING: China Airlines Flight 676 from Bali to Taipei crashes
PLACE: Taipei, Taiwan
WHEN: February 16, 1998
TOPIC EXPLICATION: The flight was from Bali to Taipei. It crashed several yards short of the runway and all 196 on board were believed dead. China Airlines had an already sketchy safety record. This crash also killed many people who lived in the residential neighborhood where the plane hit the ground. Stories on topic include any investigation into the accident, stories about the victims/their families/the survivors. Also on topic are stories about the ramifications for the airline.

Figure 2: Official description of *China Airlines crash* topic.

Relation Frequency	First Element	Second Element
0.0212	China Airlines	Taiwan
0.0191	China Airlines	Taipei
0.0170	China Airlines	Monday
0.0170	Taiwan	Monday
0.0170	Bali	Taipei
0.0148	Taipei	Taiwan
0.0148	Bali	Taiwan
0.0148	Taipei	Monday
0.0127	Bali	Monday
0.0127	International Airport	Taiwan

Table 1: Top 10 named entity pairs for the *China Airlines crash* topic.

TDT provides descriptions of each topic that annotators use to select appropriate documents by issuing and modifying IR queries. The official description of one topic ("China Airlines crash") is given in Figure 2. Table 1 shows the top 10 pairs of named entities extracted from the topic at the first stage of our algorithm (before considering connectors). The normalized relation frequency is calculated by dividing the score of the current relation (how many times we see the relation within a sentence in the topic) by the overall frequency of all relations within this topic.

It is clear from the table that the top relations mention the airline company whose plane crashed (*China Airlines*), where the crash happened (*Taiwan, Taipei, International Airport*), where the plane was flying from (*Bali*), and when

Relation	Connector	Connector Frequency
China Airlines – Taiwan	crashed/VBD	0.0312
	trying/VBG	0.0312
	burst/VBP	0.0267
	land/VB	0.0267
China Airlines – Taipei	burst/VBP	0.0331
	crashed/VBD	0.0331
	crashed/VBN	0.0198
Taipei – Taiwan	–	–

Table 2: Top connectors for three of the relations in Table 1.

First named entity	Second named entity	Connectors
China Airlines	Taiwan; Taipei	crashed/VBD
		trying/VBG
		burst/VBP
		land/VB
		killing/VBG

Table 3: Final event output for the relations of Table 1.

Connector	Frequency across topic
crashed/VBD	0.0189
burst/VBP	0.0107
trying/VBG	0.0092
land/VB	0.0079

Table 4: Top connectors across the entire *China Airlines crash* topic.

the crash happened (*Monday*). Interestingly we obtain a clique for the three elements *China Airlines*, *Taiwan*, and *Taipei*. Let us analyze the connectors for the three pairs among these three elements (Table 2). The normalized connector frequency is calculated by dividing the frequency of the current connector (how many times we see this connector for the current relation) by the overall frequency of all connectors for the current relation.

China Airlines is linked to both *Taipei* and *Taiwan*, and the lists of connectors are similar enough for our system to merge the two extracted events to one. On the other hand, there is no event connector linking *Taipei* and *Taiwan*. Our system assumes that this relationship is a static one (indeed, Taipei is the capital of Taiwan), and drops this candidate event. The final output is shown in Table 3. The connectors output by the system highlight the major event linking *China Airlines* and {*Taiwan*, *Taipei*}, that is, the crash. The importance of these connectors is also verified by calculating the relative connector frequencies for the entire topic, irrespective of the specific entities involved (Table 4).

Finally, we factor in topic specificity for the extracted events. Tables 5 shows the most and least specific named entity pairs for this topic. The less specific entries correspond to generic relationships (e.g., there are only seven week days), relationships totally independent of the topic (e.g., *Bill Hazard* reports from *Washington*), and relationships related but not limited to this topic (e.g., *China* and *Taiwan* have a long relationship separate from this crash, resulting in their mention in other topics as well). In our example, the top event of Table 3 is specific to this topic, but other events further down in the list (such as the *China–Taiwan* one) are deemed non-specific and pushed further down or removed from the output.

Relation	Specificity
China Airlines – Monday	1.0000
Taiwan – Monday	1.0000
Bali – Taipei	1.0000
Beijing – Tuesday	0.5681
Bill Hazard – Washington	0.4815
Wednesday – Monday	0.2448
Tuesday – Monday	0.1922
China – Taiwan	0.1582
CNN – New York	0.0850

Table 5: Pairs which are and are not specific for the *China Airlines crash* topic.

Event elements	Verbs	Nouns
Taiwan – passengers	killing/VBG	197/CD
	carried/VDB	196/CD
		182/CD

Table 6: Event extracted for the noun "passengers" from the *China Airlines crash* topic.

We close this section with a comment on the anchor points used by our algorithm. Such anchor points (by default named entities) are necessary in order to limit the amount of relations considered. We chose named entities on the basis of our analysis of events marked by people (Section 3). However, the system is adaptable and the user can specify additional words or phrases that should be used as anchor points. In this example, it makes sense to extract information involving the passengers of the plane. If the word *passengers* is submitted to the system, then the third from the top events extracted will refer to the deaths of the passengers, as shown in Table 6.[4]

6 Comparison with Information Extraction

We compare our system's output to ideal output for one of the most well-known information extraction competitions, the Message Understanding Conferences (MUC) organized by NIST between 1992 and 1998. In MUC's Scenario Template task events are extracted for several pre-specified domains (MUC, 1997). For each domain a list of templates is created in advance and event extraction is equated to filling these templates, a typical approach in information extraction (Riloff, 1996; Grishman, 1997). Events are extracted from one text at a time and not a collection of texts.[5]. Each

[4] 197 is the correct number and it was used more often than the other two numbers which were given in the early articles describing this crash when the exact numbers were not clear yet.

[5] There are IE systems which try to fill predefined templates from several texts but during the MUC competition systems ana-

Bogota, 5 APR 90 (EFE) — Authorities reported today that liberal senator Federico Estrada Velez, 54, one of the main leaders of the ruling liberal party, was released today in Medellin by the drug trafficking organization known as the Extraditables.
Senator Estrada Velez was kidnapped on 27 March near his home by the Extraditables, the Medellin Cartel's armed wing.

Figure 3: MUC-7 text for kidnapping/release event.

Location: Colombia
Date: 27 MAR 90
Type: kidnapping
Organization: "Medellin Cartel" / "Drug Trafficking Organization" / "The Extraditables" / "The Medellin Cartel's Armed Wing"
Target: "Federico Estrada Velez"
Effect of incident: Regained freedom: "Federico Estrada Velez"

Figure 4: Ideal MUC-7 output for the article of Figure 3, using the *Terrorism* event template.

Event elements	Verbs	Nouns
Federico Estrada Velez Medellin	released/VBN	today/NN 54/CD
Estrada Velez 27 March Medellin Cartel armed wing	kidnapped/VBN	–

Table 7: Output from our system for the text of Figure 3.

text can contain one, several, or no events. The best systems achieved performance of 51% in F-measure[6] in the last MUC-7 competition (1998); the highest F-measure result was reported during MUC-4 (1992) at around 57%.

Given the short article of Figure 3, the ideal output is the template shown in Figure 4. Because each MUC template covers a single event, the model output mixes in this case information about two atomic events: Mr. Velez's kidnapping and his release. Note that as a result it is impossible to tell to say if "27 MAR 90" in the output stands for the date of kidnapping or release.

Our system produces output which specifies both events separately: release and kidnapping. According to this output (Table 7) it is possible to figure out who was the main subject of both events (*Federico Estrada Velez*), what organization kidnapped him (*Medellin Cartel*), when he was kidnapped (*27 March*), and when he was released (*today*).[7]

There are texts in MUC collection for which no templates match and therefore no events should be extracted. Here is part of such a text:

lyzed and extracted events for one text at a time.

[6]The harmonic mean of precision and recall, i.e., $2PR/(P+R)$.

[7]The last reference requires using the date of the article (*5 APR 90*) to resolve it, a capability that our system does not yet have.

Event elements	Verbs	Nouns
Virgilio Barco Francois Mitterrand Wednesday	briefed/VB	–

Table 8: Event extracted by our system from the Barco-Mitterand passage.

> *Paris (France), 5 April 90 (AFR) – Colombian leader Virgilio Barco briefed French president Francois Mitterand here Wednesday on the efforts made by Bogota to fight the county's powerful cocain traffickers. Mr. Barco told reporters after the meeting at the Elysee Palace that the French leader, who visited Bogota in October 1989, had said once again that he was "very interested" in the drug problem.*

This text is from the terrorism domain collection. And though really no terrorist attacks are described in this text it does not mean that there are no events described. These events include the meeting between François Mitterand and Virgilio Barco, Mitterand's earlier visit to Bogota, and Barco's speaking to reporters.

Though MUC systems are not supposed to output any events for the whole text from which the above passage was excerpted, our system outputs several events; Table 8 shows the extracted event that corresponds to the above passage.

In fairness to the MUC systems we note that they perform additional tasks such as the semantic classification of the information (deciding which slot to select for a given piece of extracted text). Our approach does not assign labels such as *perpetrator* or *target* to named entities. It provides for a more superficial "understanding" of the elements of the event and the roles they play in it, in exchange for increased portability, generality, and robustness.

7 System Evaluation

7.1 Methodology

To evaluate our system we chose randomly 9 topics out of the 70 TDT-2 topics containing more than 5 and less than 500 texts (see Section 5). For each of these topics we randomly chose 10 texts, and ran our system on these 10 texts only, producing a ranked list of events with verb and noun labels, as described in Section 4. We then gave the texts and the top 10 events in the system output for a given topic to a volunteer evaluator (a graduate student in computational linguistics). Each evaluator processed exactly one topic.

Our subjects were asked to first read the texts[8] and then provide a numerical score for the system in the following areas:

1. Whether the named entities in the events extracted by our system are really related to each other in the texts.

[8]Which was the reason we limited the number of texts per topic to 10.

Question	Average rating	Percentage non-zero	Percentage above 0.5
Link quality	0.7506	92.22%	74.44%
Importance	0.6793	95.00%	62.87%
Label quality	0.6178	90.91%	51.09%

Table 9: Evaluation scores for our system. Importance and label quality are measured only on extracted relations of reasonable quality (with link quality score above 0.5, 74.44% of the total extracted events).

A separate score between 0 and 1 was given for each extracted event.

2. Whether the extracted relations between named entities, if valid, are also important. Again a 0 to 1 score was assigned to each extracted event.

3. Whether the labels provided for a (valid) event adequately describe the relationship between the named entities.

For these three questions, the evaluators gave a separate score for each extracted event. Rather than "yes"/"no" answers, they were free to use a scale of their own choosing between 0 (utter failure) and 1 (complete success).

In addition, we asked evaluators to enumerate important events that the system missed, and provide a subjective rating between 0 and 1 on how closely related the articles in their set were.

7.2 Results

Table 9 shows the scores obtained during the evaluation. We report the average rating our system obtained on each of the three questions, across both the ten extracted events in each set and the nine evaluators/topics. We also report the percentage of extracted events that received a non-zero score and a score above 0.5.

We note that the easiest task for the system is to find valid relationships between named entities, where we obtain about 75% precision by either the average score or the number of scores above 0.5. Next comes the task of selecting important links, with precision of 63–68%. The hardest task is to provide meaningful labels for the events; we succeed in this task slightly in more than half of the valid extracted events, or approximately 40% of the total extracted events.

Since the difficulty of the task is correlated with the coherence of the document sets being analyzed, we observed significant differences in the scores between topics. In some cases our system obtained scores above 70% or 80% in all three questions. In two cases, the scores were below 20%; in one of those, the documents covered a very wide range of events (many different events related to the Israeli-Palestinian peace negotiations), while the other topic dealt with an earthquake in Afghanistan. In that latter case, our system looking by default for named entities could not extract enough events as no named participants were mentioned. Regardless, our system overall extracted at least somewhat useful information, as manifested by the fact that over 90% of the reported events received non-zero scores.

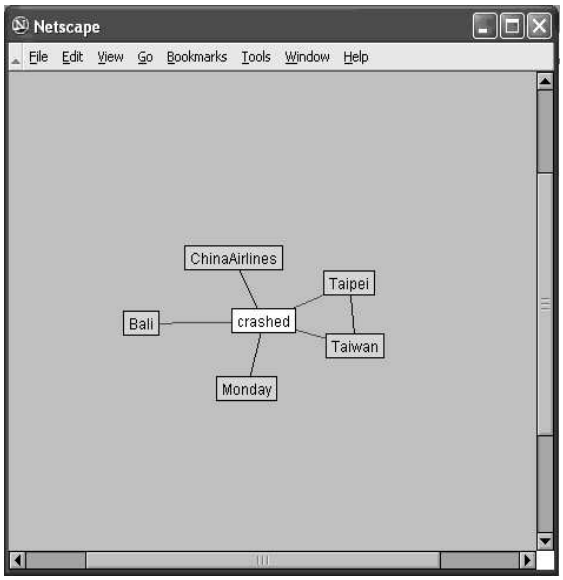

Figure 5: Screen shot of the system's output visualizing the entities connected with the verb "crash" in the *China Airlines crash* topic.

8 Conclusions and Future Work

We have reported on an empirical study of event annotation and a new approach for detecting events in text that draws from disparate earlier approaches in linguistics, information retrieval, and information extraction. We have implemented a robust, statistical system that detects, extracts, and labels atomic events at the sentence level without using any prior world or lexical knowledge. The system is immediately portable to new domains, and utilizes information present in similar documents to automatically prioritize events that are specific (and therefore likely more interesting) to a given set of documents. Our examination of results and a first small-scale evaluation indicate that the approach is promising as a means for obtaining a shallow interpretation of event participants and their relationships.

We believe that this approach can enable significant new techniques for a number of natural language processing tasks. A first direction is to use the extracted event information as a means of indexing the documents; rather than using keywords, we can now use the important events in each document, and index on participants, time phrases, and locations. This will allow the retrieval of related documents in which the same person or organization plays a prominent role, or which describe events in comparable time frames. The list of events itself may provide a different kind of indicative summary than the summaries currently based on extracted or reformulated sentences. We have already implemented a visualization prototype that allows a user to observe a two-dimensional representation of the important named entities in a set of documents and their labeled interrelationships. Figure 5 shows a screen capture from our

visualization interface, depicting the main event elements linked through the verb "crash" in the *China Airlines crash* topic.

Finally, this work has been motivated by our work on question answering, where we are examining directions that would take us away from the (largely successful with current technology) answering of simple factual questions. The event representation provides a way to answer questions about "who did what to whom, when and where", as we noted earlier. More importantly, we hope that it will be a useful tool in answering more difficult and abstract questions, for instance about the similarities or differences of actions by two different actors, or about the development of a series of related actions in time.

Acknowledgments

We wish to thank Kathy McKeown and Becky Passonneau for numerous comments and suggestions on earlier versios of our system, and John Chen for providing tools for preprocessing and assigning parts of speech to the text. We also thank the members of the Natural Language Group and other graduate students at Columbia University who participated in our evaluation experiments. This work was supported by ARDA under Advanced Question Answering for Intelligence (AQUAINT) project MDA908-02-C-0008. Any opinions, findings, or recommendations are those of the authors and do not necessarily reflect ARDA's views.

References

(Allan et al., 1998) James Allan, Jaime Carbonell, George Doddington, Jonathan Yamron, and Yiming Yang. Topic detection and tracking pilot study: Final report. In *Proceedings of the DARPA Broadcast News Trascription Workshop*, April 1998.

(Allan et al., 2001) James Allan, Rahul Gupta, and Vikas Khandelwal. Topic models for summarizing novelty. In *Proceedins of the Workshop on Language Modeling and Information Retrieval*, Pittsburgh, Pennsylvania, May–June 2001.

(Bach, 1986) Emmon Bach. The algebra of events. *Linguistics and Philosophy*, **9**:5–16, 1986.

(Bikel et al., 1999) D. Bikel, R. Schwartz, and R. Weischedel. An algorithm that learns what's in a name. *Machine Learning*, **34**:211–231, 1999.

(Chaniak, 2000) Eugene Charniak. A maximum-entropy-inspired parser. In *Proceedings of NAACL'00*, pages 132–139, 2000.

(Chung and Timberlake, 1985) Sandra Chung and Alan Timberlake. *Tense, Aspect, and Mood*, volume 3, chapter 4. Cambridge University Press, 1985.

(Filatova and Hovy, 2001) Elena Filatova and Eduard Hovy. Assigning time-stamps to event-clauses. In *Proceedings of the workshop on Temporal and Spatial Information Processing, ACL*, Toulouse, France, July 2001.

(Fiscus et al., 1999) Jon Fiscus, George Doddington, John Garofolo, and Alvin Martin. NIST's 1998 Topic Detection and Tracking evaluation (TDT2). In *Proceedings of the 1999 DARPA Broadcast News Workshop*, pages 19–24, Herndon, Virginia, 1999.

(Grishman 1997) Ralph Grishman. Information extraction: Techniques and challenges. In M. T. Pazienza, editor, *Proceedings of the Information Extraction International Summer School (SCIE-97)*, 1997.

(Levin, 1993) Beth Levin. *English Verb Classes and Alternations: A Preliminary Investigation*. The University of Chicago Press, 1993.

(Marsh and Perzanowski, 1997) E. Marsh and D. Perzanowski. MUC-7 evaluation of IE technology: Overview of results. In *Proceedings of the Seventh Message Understanding Conference*, 1997.

(McCoy and Strube, 1999) Kathleen F. McCoy and Michael Strube. Taking time to structure discourse: Pronoun generation beyond accessibility. In *Proceedings of the 1999 Meeting of the Cognitive Science Society*, 1999.

(Miller et al., 1990) G. Miller, R. Beckwith, C. Fellbaum, D. Gross, and K. Miller. Introduction to WordNet: An online lexical database. *International Journal of Lexicography*, **3**(4):235–312, 1990.

(MUC, 1997) Message Understanding Conference. *Proceedings of the Seventh Message Understanding Conference*, 1997.

(Pustejovsky, 2000) James Pustejovsky. *Events and the Semantics of Opposition*, pages 445–482. CSLI Publications, 2000.

(Riloff, 1996) Ellen Riloff. Automatically generating extraction patterns from untagged text. In *Proceedings of the National Conference of the American Association for Artificial Intelligence (AAAI)*, 1996.

(Siegel and McKeown, 2000) Eric V. Siegel and Kathleen R. McKeown. Learning methods to combine linguistic indicators: Improving aspectual classification and revealing linguistic insights. *Computational Linguistics*, December 2000.

(Yang et al., 1999) Yiming Yang, Jaime Carbonell, Ralf Brown, Tom Price, Brian T. Archibald, and Xin Liu. Learning approaches for detecting and tracking news events. *IEEE Intelligent Systems*, **14**(4), 1999.